12-2015

SalAd: A Multimodal Approach for Contextual Video Advertising

Chen Xiang
National University of Singapore

Tam Nguyen
University of Dayton, tnguyen1@udayton.edu

Mohan Kankanhalli
National University of Singapore

SalAds: A Multimodal Approach for Contextual Video Advertising

Chen Xiang
School Of Computing
National University of Singapore
Singapore
chxiang@comp.nus.edu.sg

Tam V. Nguyen
Department of ECE
National University of Singapore
Singapore
tamnguyen@nus.edu.sg

Mohan Kankanhalli
School Of Computing
National University of Singapore
Singapore
mohan@comp.nus.edu.sg

***Abstract*—The explosive growth of multimedia data on the internet creates huge opportunities for online video advertising. In this paper, we propose a novel advertising system called SalAds, which utilizes textual information, visual content and the webpage saliency, to automatically associate the most proper companion ads with online videos. Unlike most existing approaches that only focus on selecting the most relevant ads, SalAds further considers the saliency of selected ads to reduce intentional ignorance. SalAds consists of three basic steps. Given an online video and a set of advertisements, we first roughly identify a set of relevant advertisements based on the textual information matching. We then carefully select a set of candidates based on the visual content matching. In this regard, our selected ads are contextually relevant to online video content in terms of both textual information and visual content. We finally select the most salient ad among the relevant ads as the most appropriate one. To demonstrate the effectiveness of our method, we conduct a rigorous eye-tracking experiment on two ad-datasets. Our experimental results show that our method enhances the user engagement with the ad content, and yet maintain users' video viewing experience, when compared with existing approaches.**

***Keywords*-online video advertising, contextual relevance, saliency, eye gaze**

I. Introduction

The past decade has witnessed the tremendous growth rate of the World Wide Web, and therefore, online video sharing websites such as YouTube, Youku, and Facebook are fast becoming potential alternatives for video content generation and distribution. The explosive growth of multimedia data on the internet creates huge opportunities for multimedia advertising. As a typical example, Google snapped up YouTube at the price of $1.65 billion in 2006, while the latter in turn pulled in about $4 billion in revenue during 2014 [1]. Moreover, as estimated in [2], online video is growing faster than most other advertising formats and mediums, and video ad revenue will increase at a three-year compound annual growth rate of 19.5% through 2016. To take the maximum advantage of this increasing market share, video advertising has become a very important monetization strategy for many online media sharing companies. In this paper, we focus on companion advertising, where a banner/text/image ad is displayed by the top-right side besides the video. Associated with an online video are the multiple sources of information, including textual description, visual content, and user demographic information (such as geography information from IP address, age, gender). Based on different types of used information, we classify most existing contextual video advertising strategies into three categories: text-based advertising, visual-based advertising, and targeted-advertising.

A. Text-based Advertising

Typically, there are two ways to acquire the textual information of an online video, one may either mine the existing text (i.e., in YouTube, video authors are required to provide the title, description, keywords before they upload the videos) or obtain generated text (i.e., using Optical Character Recognition, speech recognition techniques to generate the video scripts). Particularly, Okada et al. [3] took advantage of five types of video metadata including title, description, keywords, category and comments to retrieve relevant ads without the necessity of expensive image and video processing. Lee et al. [4] extracted advertising keywords on particular scene of video content using corresponding scripts. Moreover, [5] relied on both the ancillary text (the surrounding text such as title and description of the video) and the video scripts for better ad-selection performance.

B. Visual-based Advertising

Most visual advertising approaches fall into the following scenarios: logo oriented, product oriented, actor oriented, and scene oriented. In logo/product oriented advertising, when a logo/product appears in the video, ads of the same brand/product will be associated with the corresponding shot. For example, [6] adopted SIFT features and SVM classifier to detect and recognize advertising trademarks in sports videos, [7] leveraged content-based object retrieval techniques to identify the objects in the video frame first and then determined the ads by matching image objects. In actor oriented advertising, the ads will be allocated to the video containing the same actor/actress. A representative example is the vADeo system [8] in which a face recognition system is used for recognizing actors/sports-persons in the movies and sports videos. In scene oriented advertising, the ads and video content are matched at the semantic level, for example, kitchen can be connected with food and dishes.

Dong et al. [9] considered all the above strategies. Due to the computational complexity of image and video processing, the authors only consider limited scenarios.

C. Targeted Advertising

Apart from user activities on the social platform, the available information about the user can be collected from browser's cookies, and user registration information. [10] classified users into different groups based on their age, location and gender. Then they identified the interests of user group from market data. Xu et al. [11] used eye-tracking to detect the region of interest and further applied it to online personalized document, image and video recommendation. Yadati et al. [12] developed an interactive advertising system by introducing two features: arousal (captured by eye-tracking tools) and valence (captured by facial expression analysis tools).

It is observed that most of the above works focus on contextually relevant ad-selection based on one particular domain of information. ImageSense [13] and VideoSense [14] combined both textual information and visual content for more relevant ads. Recent works [15] have discovered the existence of intentional ignorance or ad blindness that users tend to ignore the text and image located on the right side of the webpage. In this paper, we use textual information and visual information to select contextually relevant ads, and then consider the webpage saliency to reduce ad blindness.

To foster a vigorous and healthy online video advertising ecosystem, we argue that: a) the selected ad should be contextually relevant to the given video, in terms of both textual relevance and visual relevance. The textual information is the drastic summarization of the video, and visual content reflects user's attention directly. We believe that the combination of textual relevant and visual relevance will result in better selection. b) The selected ad should be salient so that users will notice it when the video is playing. On one hand, users tend to focus their attention on the video content and ignore the companion ads. On the other hand, it is obvious that users cannot always concentrate on the video content all the time. Their attention varies as the video content changes. The salient ad therefore allows for greater possibility to draw more users' attention.

Motivated by the observations above, we propose a novel video advertising system named SalAds. We highlight the contribution of our work as three-fold:

1) We introduce a novel feature for ad-selection, namely, the webpage saliency. The mainstream of contextual advertising uses textual information or visual information to select relevant advertisement(to ensure users engagement). As the best of our knowledge, we are the first to explore the webpage saliency for improving ad-selection.
2) We demonstrate the effectiveness of our method on two newly built ad datasets, and compare it with two other typical baselines.
3) We conduct experiment on a video playing environment and employ eye-tracking techniques to explore understand actual users behavior.

The rest of this paper is organized as follows: Section II describes our SalAds system. Section III presents our experiment settings and experimental results. Section IV summarizes this paper and discusses our future work.

II. The Proposed Method

A. Problem Formulation

In companion advertising, we aim at addressing the following two problems:

1) how to select relevant ads to ensure user engagement;
2) how to select attractive advertisement to reduce ad blindness.

Given a video and a set of ads, our SalAds system should allocate a proper ad to the video within seconds.

Let V denote an online video, which is associated with two typical attributes, metadata and video frames, represented by vm and vf. Let the video be represented by k keyframes $vf = \{vk_i\}, 1 \leq i \leq k$. Let A denote a set of n candidate ads $A = \{a_j\}, 1 \leq j \leq n$. Similarly, each ad a_j can be represented by metadata provided by the advertisers am_j and visual content af_j. Let W denote the webpage excluding the video frame area and the ad area. We take the rest of the webpage into consideration since user's attention is distributed across the whole webpage, and we want our selected ads to draw more attention from the user. The contextual relevance $R(V, a_j)$ between the video V and an ad a_j is given by the linear combination of textual relevance $R_{text}(vm, am_j)$ and visual relevance $R_{visual}(vf, af_j)$:

$$R(V, a_j) = \lambda_1 R_{text}(vm, am_j) + \lambda_2 R_{visual}(vf, af_j) \quad (1)$$

where λ_1 and λ_2 are the corresponding weights, and $\lambda_1 + \lambda 2 = 1$. In this way, a list of ads can be ranked according to the contextual relevance to a given video.

To reduce ad blindness, we want the inserted ads to be salient[14]. The saliency value $S(a_j)$ of ad a_j is measured by the contrast between the ad and the context:

$$S(a_j) = \beta_1 C(vf, af) + \beta_2 C(W, af) \quad (2)$$

where $\beta_1 C(vf, af)$ and $\beta_2 C(W, af)$ denote the contrast from video frames and webpages. Without loss of generality, we notice that our SalAds system can be easily extended to improve existing webpage-based contextual advertising and targeted advertising.

B. System Overview

Fig. 1 illustrates the overall system framework of SalAds. There are three major components, namely, text-based ad ranking, visual-based ad ranking and attention-based ad ranking. In text-based ad ranking, we rank the ads based on

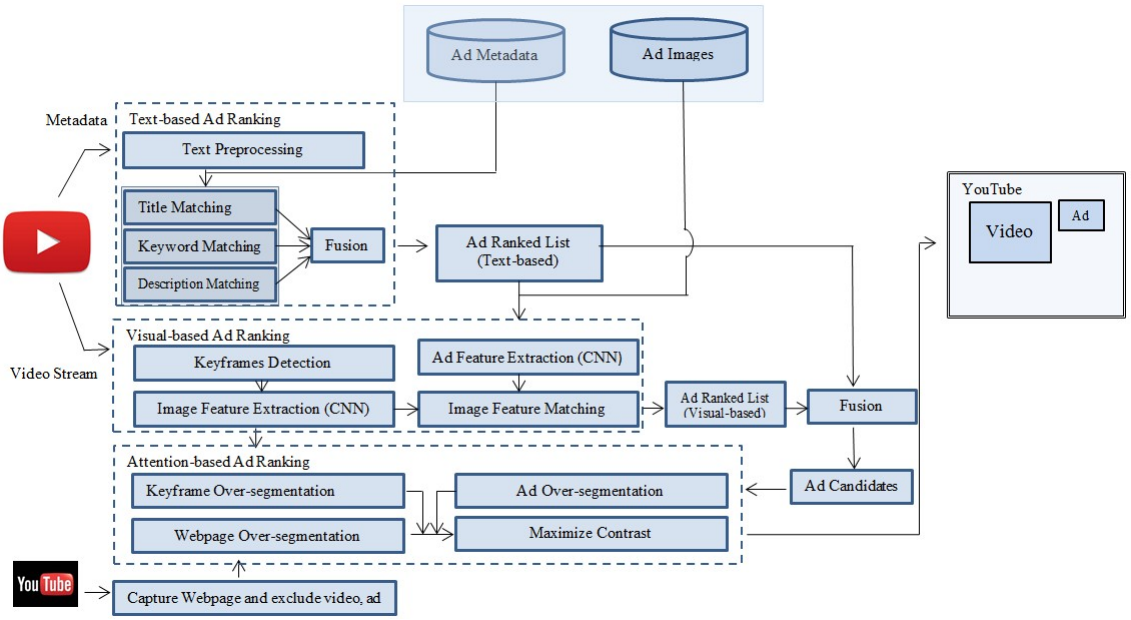


Figure 1. System framework of SalAds. The ad is inserted in the top-right area. SalAds consists of three major components, namely, text-based ad ranking, visual-based ad ranking, and attention-based ad ranking.

the title, keywords and description match between video and ads. For each type of textual information, we use the Vector Space Model (VSM) to measure the similarity, and then we use linear combination to fuse the matching results. To reduce unnecessary computation, we select a set of textual relevant ads as the candidates for visual-based ad ranking. In visual-based ad ranking, we first extract keyframes for each input video, and then extract visual features. The ads are ranked based on the visual feature matching. Then we fuse the text-based ad ranking list and the visual-based ad ranking list to select the contextually relevant ads as the candidates for attention-based ad ranking. In attention-based ad ranking, we over-segment the keyframes, webpage and ad image. The ad candidate with the maximum contrast will be selected as the proper ad.

C. Textual Relevance

The surrounded textual information of the videos provides a general summary about the video content. We adopt the Vector Space Model (VSM) to measure the textual relevance since [16] proved that matching and ranking the ads based on VSM is the best among a set of simple methods. Intuitively, each ad a_j and video V is represented as a vector of weights $a_j = (w_{j,1}, w_{j,2}, w_{j,3} \dots w_{j,T})$ and $V = (w_1, w_2, w_3 \dots w_T)$, while the dimension T of vector is the number of distinct terms in the dictionary. The similarity $R_{text}(V, a_j)$ between V and a_j is calculated by the cosine distance of the two vectors.

$$R_{text}(V, a_j) = \frac{V \cdot a_j}{\|V\| \cdot \|a_j\|} \tag{3}$$

In SalAds, we consider three types of textual information of an online video, including title, keywords and description. As for the ads, we also collect the corresponding textual information. Unlike traditional textual information matching methods, we treat these three sources of textual information separately by assigning different weights. In this regard, the similarity $R_{text}(V, a_j)$ is given by a weighting based VSM.

$$\begin{aligned} R_{text}(V, a_j) = \alpha_1 R_{title}(V, a_j) + \alpha_2 R_{keywords}(V, a_j) \\ + \alpha_3 R_{desctiption}(V, a_j) \end{aligned} \tag{4}$$

where $R_{title}(V, a_j)$, $R_{keywords}(V, a_j)$, $R_{desctiption}(V, a_j)$ are the textual relevance when using one source of text information, and α_1, α_2, α_3 indicate the contribution from corresponding component to the overall textual relevance, $\alpha_1 + \alpha_2 + \alpha_3 = 1, 0 \leq \alpha_1, \alpha_2, \alpha_3 \leq 1$. The weight of each type of textual relevance is evaluated accoding to the $Spread$ strategy described in [3]. We collect 1000 popular videos from YouTube, and information about each video includes title, description, and keywords. The corresponding value of α_1, α_2, α_3 are empirically set as 0.45, 0.30, 0.25, respectively.

D. Visual Relevance

In visual-based video advertising, identifying the products, people, and logos is a typical way to find relevant ads. In this regard, the visual-based relevance matching problem is transfered as a traditional content based image retrieval problem. In the area of image content analysis, the fundamental technique is extracting image features to represent the images. To date, the features can be classified into

the following three categories: low-level features (e.g. color histogram, HOG, and SIFT), mid-level features (e.g. spatial pyramids, bags of features, and higher layer activations of convolution neural networks) and high-level features (also known as semantic features). Because of the outstanding performance of convolution neural network (CNN) in image content analysis [17], we adopt it as our image feature extraction tool. The dimension of the image feature vector is 4096. Let vf_i denote the feature vector of the keyframe i, and af_j denote the feature vector of candidate ad a_j. The visual relevance $R_{visual}(V, a_j)$ between video V and ad a_j is given by

$$R_{visual}(V, a_j) = \max_{1 \le i \le k} \{ \frac{vf_i \cdot af_j}{\|vf_i\| \cdot \|af_j\|} \} \tag{5}$$

It is obvious that we match each candidate ad with all the k keyframes, and the maximum value is the visual-based matching score. It is reasonable that there is high relevance between video V and ad a_j if a_j matches any keyframe.

Since the scale of video-ad matching is daunting, over a course of a week, the ad-network involves billions of impressions, hundreds of millions of distinct pages, and hundreds of millions of ads. Unsurprisingly, given an online video, the vast majority of ads are irrelevant. To reduce the unnecessary cost of image feature extraction and matching, we use the text-based ad ranking procedure to narrow down the search space.

E. Attention Ranking

We define the attention as the joint saliency of the webpage and the candidate ad. To reduce intentional ignorance of companion ads, we propose to select the salient ads in the context. It is observed that users cannot always concentrate on the video content all the time, and their attention wonders as the video content changes. The salient ad has greater capacity to draw more users attention. As [14] indicates that the "positive" relevance ensures higher similarity between video and ad content, while the "negative" relevance will gain more attention because of high contrast. In this regard, the ads with high contrast to the video content and the rest of the webpage will be the ideal choice. To address this problem, we adopt the simple-linear-iterative-clustering (SLIC) [18] to parse the video frame (denoted by vf_i), the ad images (denoted by af_j), and the rest of webpage (denoted by w), into superpixels. Noticing that the left and right border of YouTube are blank where users tend to pay little attention to. In our work, w capture the webpage excluding the video frame area, ad area, and the blank area on the left and right border. Each superpixel is represented by the mean (L, A, B) of all pixels within the superpixel. We choose LAB color space because it is fast, and color is a good cue to attract human attention [18]. Based on the size of corresponding area, we set the number of superpixels for video frame, ad image, and the rest of webpage as $nv(= 50)$, $na(= 20)$, $nw(= 100)$. Thus, each area is represented by a set of 3-dimensional vectors:

$$\begin{aligned} vf_i &= (l_{i,p}, a_{i,p}, b_{i,p}), 1 \le p \le nv \\ af_j &= (l_{j,q}, a_{j,q}, b_{j,q}), 1 \le q \le na \\ w &= (l_r, a_r, b_r), 1 \le r \le nw \end{aligned} \tag{6}$$

The contrast $C(vf, af_j)$ between video v and ad a_j is given by

$$C(vf, af_j) = \max_{1 \le i \le k} \{ \frac{1}{nv \times na} \times \left(\sum_{q=1}^{na} \sum_{p=1}^{nv} \|vf_{i,p} - af_{j,q}\|_2 \right) \} \tag{7}$$

The contrast $C(w, af_j)$ between the rest of webpage w and ad a_j is given by

$$C(w, af_j) = \frac{1}{nw \times na} \times \left(\sum_{q=1}^{na} \sum_{r=1}^{nw} \|w_r - af_{j,q}\|_2 \right) \tag{8}$$

Since users tend to pay more attention on the video area and less attention on the rest of webpage, it is reasonable to assign a larger weight to $C(vf, af_j)$ and a smaller weight to $C(w, af_j)$. In our experiment, we empirically set $\beta_1 = 0.8$ and $\beta_2 = 0.2$. Thus, the contextually relevant ads can be further ranked according to (2), (7), (8). The ads with the maximum contrast value will be selected as the proper ads.

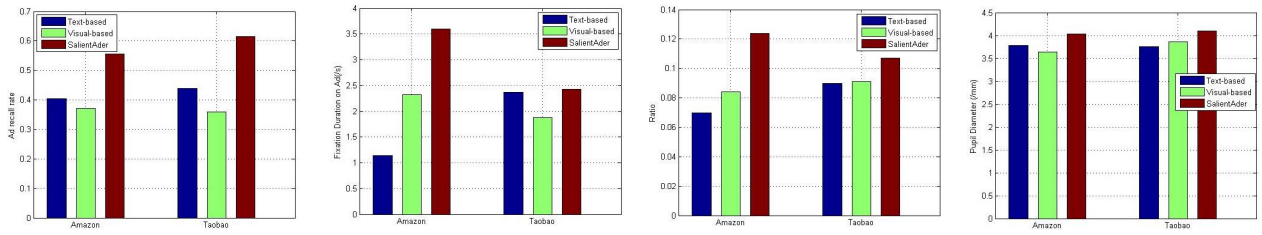
III. Experiments and Evaluations

A. Data Collection

We collected 1046 popular YouTube videos to construct our video dataset. For each downloaded video, we crawled the textual information including title, keywords and description. According to the statistics from Sysomos[1], we chose the top 10 most popular categories and for each category, we selected 2 most viewed videos. We took the selected 20 videos for evaluation. The average viewing rate of these videos is over 100 million. The average length of these videos is around 3.5 minutes (according to YouTube Charts[2], the majority of top 500 most viewed videos are short videos).

In order to avoid language bias, we tested our system on two separate ad-datasets, one of English ad-dataset and another one of Chinese ad-dataset. We utilized the products from Amazon[3] and Taobao[4] for compiling our ad-dataset. Our Amazon ad-dataset consists of 93424 ads and Taobao ad-dataset consists of 140532 ads. For each ad, we downloaded the ad image and the textual information including title, keywords, and description. We disregard the category since the vocabulary of video and ads is quite different, and category of online videos is usually represented by one item.

[1] http://sysomos.com/reports/youtube-video-statistics
[2] http://www.youtube.com/charts
[3] http://www.amazon.com/
[4] https://www.taobao.com/



(a) (b) (c) (d)

Figure 2. User engagement comparison on the two ad datasets. From left to right, (a) compare ad recall rate, (b) compare fixation duration on ads, (c) compare ratio, (d) compare pupil diameter on ads (please view in high 200% resolution).

B. Experiment Setting

We conduct a thorough and systematic eye-tracking experiment to test the efficiency of our method, when compared with two other baselines, text-based advertising method and visual-based advertising method. We used a binocular infrared based remote eye-tracker SMI RED 250 to record users' eye gaze data, and we also employ an immediate cued recall (for each displayed ad, we mixed it with four other similar ads, and the user was supposed to intuitively choose the ads s/he has seen) to assess users' memory about the displayed ads. We invited 60 volunteers to participate in the experiment, and all of them are from the university in the age-group of 20-30. To simulate a real-world online video viewing environment, we built a browser interface which is similar to YouTube. Each user was shown 10 videos with selected companion ads. The users had no prior knowledge about the purpose of the experiment, and they were told to watch the videos as they did on their own device. At the end of the experiment, users were asked to recall the ads.

We divided the 60 volunteers into two groups according to their language preferences: 30 volunteers viewed videos with English ads, and the other 30 viewed videos with Chinese ads. For each ad-dataset, each advertising strategy was assigned to 10 volunteers. Thus, we have 100 samples of data which is sufficient for analyzing the average score. In case of any kind of bias, we made sure that each video within each ad-dataset and each strategy has been viewed for 5 times, and the sequence of displaying 10 videos is randomized.

C. Experimental Results

1) Ad Recall: The ad recall rate is the most direct manner to measure how much the users assimilate the ad content Fig. 2 (a) presents the results of ad recall rate comparison on two ad datasets, where Y-axis represents the average ad recall rate. We observe that in both ad-datasets, the recall rate of our method is above 0.55, while the recall rate of text-based and visual-based ad ranking strategy is around 0.4, which indicates the outstanding performance of our method. Our user-study experiment provides evidence about the existence of blindness in companion video advertising, and we believe that if we do uncued-recall, the value will be even lower. Our method with considering the saliency of the webpage has a significant improvement. This is not surprising given that: usually, the companion ads are quite simple, and it takes a short time for the users to read and further remember them. However, it cannot ensure a 100% recall rate especially when the users are completely immersed in the video content.

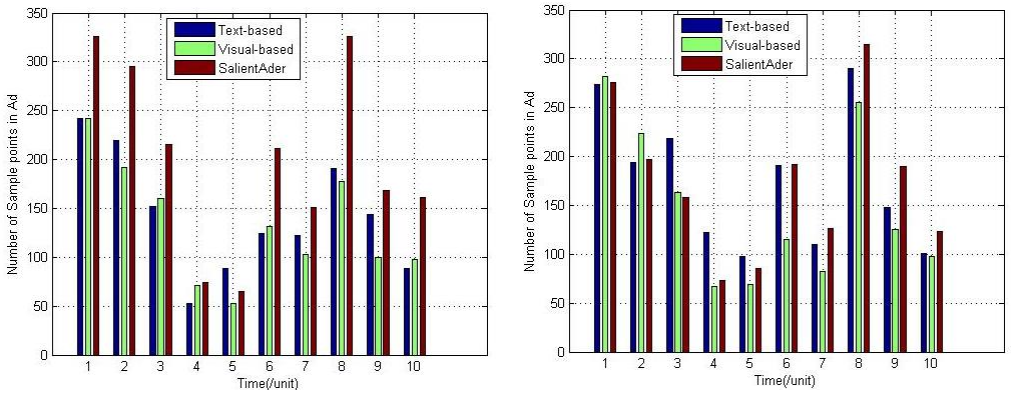
2) User Engagement: We analyzed the recorded data from two aspects: event-based and sample-point-based. In eye-tracking techniques, there are three typical events, namely, saccade, fixation and blink. Since user does not gain information from the outside world from saccades and blinks, we only consider the fixations. Intuitively, the accumulative fixation time on the ad area indicates user's attention on the ads. Usually, the longer the fixation time is, the better the user will remember the ads. As illustrated in Fig. 2 (b), our method arouses more attention than the baseline methods.

The sample-point-based analysis is to count the number of points where the users' eye-gaze fixated during the video playing. It is obvious that users tend to concentrate on video content. When they divert their attention from the video, hopefully we want them to assimilate the ads more then the rest of webpage. In this regard, we define the ratio as

$$ratio = \frac{\#sample\ points\ in\ ad\ area}{\#sample\ points\ in\ rest\ webpage} \tag{9}$$

In our experiment, we recorded eye-movement data at a frequency of 250HZ. The normal size of YouTube video frame is (850×310), the size of companion ad is (400×300), and the webpage size is (1920×1080). If a user's attention is uniformly distributed, the ratio is 0.07. As illustrated in Fig. 2 (c), we observe that the average ratio under our method is above 0.10 (greater than 0.07), while the ratio under other two method is around 0.08 (slightly greater than 0.07), which indicates the significant improvement of our method.

We also evaluate the pupillary dilation since it measures user's interest and engagement levels [19]. In Fig. 2 (d), we plot the average pupillary dilation of the sample points on ads. The average pupillary dilation of our method is larger than the average pupillary dilation of the baselines.



(a) (b)

Figure 3. User comparison comparison on the two ad datasets. (a) Comparison on Amazon ad-dataset. (b) Comparison on Taobao ad-dataset.

Our method induces the highest user engagement.

3) User Intrusiveness: We measure the intrusiveness objectively through figuring out at which point the users tend to view the ads. Because the selected videos have various lengths, the longer videos necessarily have larger number of points. To avoid this bias, we normalize all the videos into 10 units, and for each unit, we count the number of samples on ads. As illustrated in Fig. 3, for all the three methods, users tend to view the companion ads at the prelude and epilogue since most eye-gaze points fall within the first and last three units. What is more, our method arouses more user attention during the two period in both ad-datasets. While in the middle of video playing, our method does not arouse extra intrusiveness when compared to the two baselines.

IV. Conclusions

In this paper, we have presented a novel and effective contextual video advertising system called SalAds. We combined textual information and visual content to select relevant ad candidates, and further consider the saliency of webpage to select the most proper ad. Through a thorough user-study and eye-tracking experiment, we demonstrate that SalAds enhances the user engagement with the ad content, and yet maintains users' online video viewing experience. We consider the saliency in terms of color contrast in this paper, we may integrate motion saliency into SalAds in our future work. For example, since we notice that users tend to view ads at the prelude and epilogue, a rotating ad appearing at the beginning or fading out at the end for a certain duration is likely to arouse user attention.

References

[1] Silicon Beat. http://www.siliconbeat.com/.

[2] Business Insider. http://www.businessinsider.com/.

[3] K. Okada, E. S. de Moura, M. Cristo, D. Fernandes, M. A. Gonçalves, and K. Berlt, "Advertisement selection for online videos," in *Symposium on Multimedia and the web*. ACM, 2012, pp. 367–374.

[4] T. Lee, Jung, H. Lee, H.-S. Park, Y.-I. Song, and H.-C. Rim, "Finding advertising keywords on video scripts," in *ACM Conference on Research and development in information retrieval*. ACM, 2009, pp. 686–687.

[5] T. Mei, J. Guo, X.-S. Hua, and F. Liu, "Adon: Toward contextual overlay in-video advertising," *Multimedia systems*, vol. 16, no. 4-5, pp. 335–344, 2010.

[6] L. Ballan, M. Bertini, and A. Jain, "A system for automatic detection and recognition of advertising trademarks in sports videos," in *ACM Multimedia*. ACM, 2008, pp. 991–992.

[7] J. Hu, G. Li, Z. Lu, J. Xiao, and R. Hong, "Videoader: a video advertising system based on intelligent analysis of visual content," in *Conference on Internet Multimedia Computing and Service*. ACM, 2011, pp. 30–33.

[8] S. H. Sengamedu, N. Sawant, and S. Wadhwa, "vadeo: video advertising system," in *Proceedings of the 15th international conference on Multimedia*. ACM, 2007, pp. 455–456.

[9] C. Dong, S. Chen, and X. Tang, "Advisual: a visual-based advertising system," in *ACM Multimedia*. ACM, 2013, pp. 433–434.

[10] H. S. Neshat and M. Hefeeda, "Smartad: A smart system for effective advertising in online videos," in *International Conference on Multimedia and Expo*. IEEE, 2011, pp. 1–6.

[11] S. Xu, H. Jiang, and F. Lau, "Personalized online document, image and video recommendation via commodity eye-tracking," in *ACM conference on Recommender systems*. ACM, 2008, pp. 83–90.

[12] K. Yadati, H. Katti, and M. Kankanhalli, "Interactive video advertising: A multimodal affective approach," in *Advances in Multimedia Modeling*. Springer, 2013, pp. 106–117.

[13] T. Mei, L. Li, X.-S. Hua, and S. Li, "Imagesense: towards contextual image advertising," *ACM Transactions on Multimedia Computing, Communications, and Applications*, vol. 8, no. 1, p. 6, 2012.

[14] T. Mei, X.-S. Hua, L. Yang, and S. Li, "Videosense: towards effective online video advertising," in *ACM Multimedia*. ACM, 2007, pp. 1075–1084.

[15] J. W. Owens, B. S. Chaparro, and E. M. Palmer, "Text advertising blindness: the new banner blindness?" *Journal of Usability Studies*, vol. 6, no. 3, pp. 172–197, 2011.

[16] B. Ribeiro-Neto, M. Cristo, P. B. Golgher, and E. Silva de Moura, "Impedance coupling in content-targeted advertising," in *ACM conference on Research and development in information retrieval*. ACM, 2005, pp. 496–503.

[17] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.

[18] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Susstrunk, "Slic superpixels compared to state-of-the-art superpixel methods," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 34, no. 11, pp. 2274–2282, 2012.

[19] H. Katti, K. Yadati, M. Kankanhalli, and C. Tat-Seng, "Affective video summarization and story board generation using pupillary dilation and eye gaze," in *Multimedia (ISM), 2011 IEEE International Symposium on*. IEEE, 2011, pp. 319–326.